MapReduce in Beam (Python) 2.5

## Overview

In this lab, you will identify Map and Reduce operations, execute the pipeline, use command line parameters.

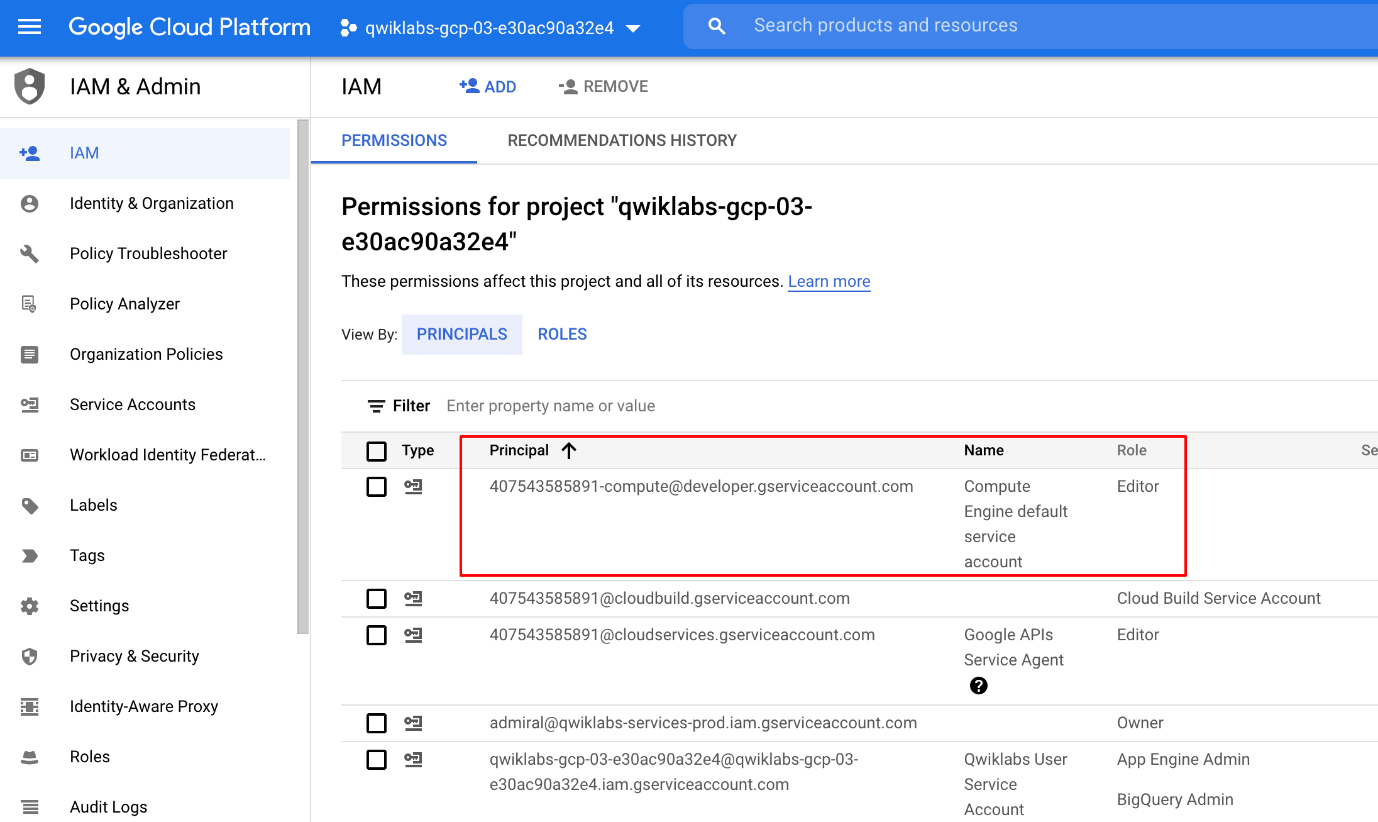
## Objective

* Identify Map and Reduce operations
* Execute the pipeline
* Use command line parameters

### **Check project permissions**

Before you begin your work on Google Cloud, you need to ensure that your project has the correct permissions within Identity and Access Management (IAM).

1. In the Google Cloud console, on the **Navigation menu** (Navigation menu icon), click **IAM & Admin** > **IAM**.
2. Confirm that the default compute Service Account {project-number}-compute@developer.gserviceaccount.com is present and has the editor role assigned. The account prefix is the project number, which you can find on **Navigation menu** > **Home**.



If the account is not present in IAM or does not have the editor role, follow the steps below to assign the required role.

* In the Google Cloud console, on the **Navigation menu**, click **Home**.
* Copy the project number (e.g. 729328892908).
* On the **Navigation menu**, click **IAM & Admin** > **IAM**.
* At the top of the **IAM** page, click **Add**.
* For **New principals**, type:

{project-number}-compute@developer.gserviceaccount.com

Copied!

content\_copy

Replace {project-number} with your project number.

* For **Role**, select **Project** (or Basic) > **Editor**. Click **Save**.

## Task 1. Lab Preparations

Specific steps must be completed to successfully execute this lab:

### **Open the SSH terminal and connect to the training VM**

You will be running all code from a curated training VM.

1. In the Console, on the **Navigation menu** (Navigation menu icon), click **Compute Engine** > **VM instances**.
2. Locate the line with the instance called **training-vm**.
3. On the far right, under **Connect**, click on **SSH** to open a terminal window.
4. In this lab, you will enter CLI commands on the **training-vm**.

### **Clone the training github repository**

In the **training-vm** SSH terminal enter the following command:

git clone https://github.com/GoogleCloudPlatform/training-data-analyst

Copied!

content\_copy

## Task 2. Identify Map and Reduce operations

1. Return to the **training-vm** SSH terminal and navigate to the directory /training-data-analyst/courses/data\_analysis/lab2/python and view the file is\_popular.py with Nano. **Do not make any changes to the code.** Press **Ctrl+X** to exit Nano.

cd ~/training-data-analyst/courses/data\_analysis/lab2/python

nano is\_popular.py

Copied!

content\_copy

Can you answer these questions about the file is\_popular.py?

* What custom arguments are defined?
* What is the default output prefix?
* How is the variable output\_prefix in main() set?
* How are the pipeline arguments such as --runner set?
* What are the key steps in the pipeline?
* Which of these steps happen in parallel?
* Which of these steps are aggregations?

## Task 3. Execute the pipeline

1. In the **training-vm** SSH terminal, run the pipeline locally:

python3 ./is\_popular.py

Copied!

content\_copy

1. Identify the output file. It should be **output**<suffix> and could be a sharded file.

ls -al /tmp

Copied!

content\_copy

1. Examine the output file, replacing '-\*' with the appropriate suffix.

cat /tmp/output-\*

Copied!

content\_copy

## Task 4. Use command line parameters

1. In the **training-vm** SSH terminal, change the output prefix from the default value:

python3 ./is\_popular.py --output\_prefix=/tmp/myoutput

Copied!

content\_copy

1. What will be the name of the new file that is written out?
2. Note that we now have a new file in the **/tmp** directory:

ls -lrt /tmp/myoutput\*

Copied!

content\_copy

## End your lab